This article has been accepted for publication in BJO following peer review.

The definitive copyedited, typeset version is available online at 10.1136/bjophthalmol

Title Page

Title: Development and evaluation of a deep learning model for the detection of multiple fundus diseases based on color fundus photography

Authors list:

Bing Li[1] MD, Huan Chen[1] MD, Bilei Zhang[1] MD, Mingzhen Yuan[1] MD, Xuemin Jin[2,3] MD, PhD. Bo Lei[2] MD, PhD, Jie Xu[4] MD, PhD, Wei Gu[5] MD, David Chuen Soong Wong[6] MA, PhD, Xixi He[7], Hao Wang[7], Dayong Ding[7] PhD, Xirong Li[8] PhD, Weihong Yu[1] MD, PhD*, Youxin Chen[1] MD, PhD*

*Joint corresponding authors: Youxin Chen and Weihong Yu contributed equally to the study

1. Department of Ophthalmology, Peking Union Medical College Hospital, Key Lab of Ocular Fundus Diseases, Chinese Academy of Medical Sciences Beijing China

2. Henan Provincial Peoples' Hospital, People's Hospital of Zhengzhou University, Henan Eye Institute, Henan Eye Hospital, Zhengzhou, Henan
3. Department of Ophthalmology, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China
4. Beijing Institute of Ophthalmology, Beijing Tongren Eye Center, Beijing Tongren Hospital, Capital Medical University, Beijing Ophthalmology and Visual Science Key Lab, Beijing, China
5. Beijing Aier Intech Eye Hospital, Beijing, China
6. University of Cambridge School of Clinical Medicine, UK
7. Vistel AI Lab, Visionary Intelligence Ltd., Beijing, China
8. MoE Key Lab of DEKE, Renmin University of China, Beijing, China

Correspondence:

Youxin Chen, MD, PhD, Key Lab of Ocular Fundus Diseases, Department of Ophthalmology, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences, Beijing, 100730, People’s Republic of China. E-mail:

chenyx@pumch.cn

Word count: 3185

Abstract

Aim: To explore and evaluate an appropriate deep learning system (DLS) for the detection of 12 major fundus diseases using color fundus photograph (CFP).

Methods: Diagnostic performance of a DLS were tested on the detection of normal fundus and 12 major fundus diseases including referable diabetic retinopathy (DR), pathologic myopic (PM) retinal degeneration, retinal vein occlusion (RVO), retinitis pigmentosa (RP), retinal detachment (RD), wet and dry age-related macular degeneration (AMD), epiretinal membrane (ERM), macula hole (MH), possible glaucomatous optic neuropathy (GON), papilledema and optic nerve atrophy. The DLS was developed with 56738 images and tested with 8176 images from on one internal test set and two external test sets. The comparison with human doctors were also conducted.

Results: The AUCs of the DLS on the internal test set and the two external test sets were 0.950 (95%CI, 0.942~0.957) to 0.996 (95%CI, 0.994~0.998), 0.931 (95%CI, 0.923~0.939) to 1.000 (95%CI, 0.999~1.000) and 0.934 (95%CI, 0.929~0.938) to 1.000 (95%CI, 0.999~1.000), with sensitivities of 80.4% (95%CI, 79.1%~81.6%) to

97.3% (95%CI, 96.7%~97.8%), 64.6% (95%CI, 63.0%~66.1%) to 100% (95%CI, 100%~100%), and 68.0% (95%CI, 67.1%~68.9%) to 100% (95%CI, 100%~100%) respectively, and specificities of 89.7% (95%CI, 88.8%~90.7%) to 98.1% (95%CI, 97.7%~98.6%), 78.7%(95%CI, 77.4%~80.0%) to 99.6%(95%CI, 99.4%~99.8%) and 88.1% (95%CI, 87.4%~88.7%) to 98.7% (95%CI, 98.5%~99.0%). When compared with human doctors, the DLS obtained a higher diagnostic sensitivity but lower specificity.

Conclusion: The proposed DLS is effective in diagnosing normal fundus and 12 major fundus diseases, and thus has much potential for fundus diseases screening in the real world.

Key words: Deep learning, color fundus photography, automatic detection, multiple fundus diseases

Introduction

Color fundus photography (CFP) plays an important role in detecting prevalent vision-threatening fundus diseases, such as diabetic retinopathy (DR), retinal vein occlusion (RVO), age-related macular degeneration (AMD), and glaucoma. According to recent epidemiological studies, approximately 79.6 million people worldwide will have glaucoma by 2020[1], while the number of people with AMD is expected to reach around 200 million[2]. The prevalence of diabetes around the world will reach 592 million people by 2035[3], with one-third affected by DR[4,5]. However, medical services are extremely limited worldwide. For example, in mainland China the ophthalmic human resource at the country level was only 0.14 per thousand people according to a survey in 2014[6]. This serious situation imposed a substantial burden on the large-scale screening of multiple fundus diseases for early detection.

Deep learning system (DLS)-based diagnosing and grading in ophthalmology has progressed rapidly in many conditions, including cataracts[7,8], DR[9-11], glaucoma[12], retinopathy of prematurity (ROP)[13,14], AMD[15,16], and macular telangiectasia (MacTel) type 2[17,18]. However, current studies mostly focus on one or

only a few (less than five) diseases[19,20]. To the best of our knowledge, there are still lack of efficient DL models for multiple diseases (especially more than 10) recognition using CFPs. We attribute this absence to two factors: the difficulties of establishing a large-scale multi-disease dataset for training and validation, and the technical challenges of developing a DLS suited not only for separating abnormal and normal CFPs but also for distinguishing one disease from many others.

Recently, Son et al[21] proposed a DLS for the detection of 12 major fundus abnormalities using 12 binary classification models, which could help greatly on the detection of retinal lesions. However, for diseases recognition, it still needs professional interpretation, which may bring obstacles for screening and AI assisted diagnosis if there's no trained ophthalmologists available. Also, the application of a panel of binary classification models will take much more time and computer resources than a single multiclassification model. This paper aims to develop an automated screening DLS for multiple major fundus diseases, which could be of great significance for clinical practice in the future.

Methods

The current study complied with the Declaration of Helsinki and was approved by the Ethics committee of Peking Union Medical Collage Hospital (NO. S-K631). The review board waived the need to obtain informed patient consent because of the retrospective study design and the use of fully anonymized CFPs.

Image acquisition and datasets

The selection of diseases was decided according to their prevalence and morbidity, also taking into account their clinical potential for screening using CFPs. Hence, in addition to normal fundus images we selected 12 major fundus diseases including nine retina diseases: referable diabetic retinopathy (DR), pathologic myopic (PM) retinal degeneration, retinal vein occlusion (RVO), retinitis pigmentosa (RP), retinal detachment (RD), wet and dry age-related macular degeneration (AMD), epiretinal membrane (ERM) and macula hole (MH), and three optic nerve disorders: possible glaucomatous optic neuropathy (GON), papilledema and optic nerve atrophy. The imaging diagnosis were made upon standard diagnostic criteria (eTable1). Although dry and wet AMD can be considered as the same disease of different stages[22], we still classified them into two categories considering their potential difference on

treatments and prognoses.

Since there were no publicly available datasets for the detection of multiple fundus diseases, we acquired and annotated a dataset for the development and internal test of the DLS. To test the generalizability of the model, we also collected CFPs from an independent tertiary medical center forming the external test set A and three primary hospitals forming the external test set B.

Development set: A total of 56738 CFPs taken between January 2014 and December 2018 were collected from three participating centers (Henan Provincial Peoples' Hospital, Zhengzhou, Henan, Beijing Tongren Hospital, Beijing and Beijing Aier Intech Eye Hospital, Beijing). These images formed the development dataset for the models' training and validation.

Test sets: Another 8176 CFPs were collected for the DLS testing. Among them, 3579 were from the same source of the development set and ensure the sample size of each disease reached over 100, forming the internal test set. Another 1245 CFPs from 757 patients were collected from another independent tertiary medical center (Peking Union Medical College Hospital) from 1st January 2019 to 30th June 2019, as the

external test set A. The last 3352 CFPs from 2558 patients were collected from three primary hospitals from 4th July 2017 to 14th September 2020, as the external test set B.

For each patient enrolled, only one image of each eye could be included. The detailed inclusion and exclusion criteria are provided in the online material.

After preprocessing and desensitization, the development data set was separated into a training set and a validation set with the ratio of 4:1 according to the patients' number, which means the bilateral CFPs of the same patient were assigned together to either the training set or validation set. This process was organized randomly. The three test sets were maintained independently to test the performance and generalization of the DLS.

Online annotation was carried out to label the images as normal fundus or the 12 selected diseases. A total of 17 senior board-certified ophthalmologists (with five to 12 years of experience) were randomly assigned for image annotation. Thirteen of them were assigned to label the development dataset and internal test set. The other

four doctors were assigned to label the external test sets. Images in the test sets were labeled three times by different ophthalmologists to obtain high reliability. Consistent labels by all three doctors were retained. If the label was only agreed by two doctors, then the final decision would be made by a fourth, more senior ophthalmologists (with over 10 years of experience). Images with no consistent labels or those annotated with poor quality, such as loss of focus, misalignment, excessive brightness or dimness, were excluded.

Development of evaluation of the DLS

The DLS was designed using the convolutional neural network (CNN) of SeResNext50[23] network as a multilabel model selected from four candidate CNNs with two parallel branches at the fully connected layer, one for the distinguish of normal and abnormalities and the other for the recognition of diseases it predicted to have, which could be more than one kind of diseases, simultaneously. The details are available on online materials (eFigure1).

The performance of the DLS were evaluated on the three test sets. We used the area under the receiver operating characteristic (ROC) curve (AUC), sensitivity and

specificity for assessments. The metrics were calculated for each label instead of each image, since one image could be annotated with more than one label. Information learned in our automated method was visualized for further clinical review using Class Activation Map[24], which is a CNN's visualization technique that can identify the importance of the image regions by projecting back the weights of the classification layer on the convolutional feature maps obtained from the last convolution layer.

The comparison of the DLS with human doctors

To assess if the DLS has reached a comparable diagnostic performance with human doctors, four ophthalmic residents were tested using the external test set B. They were assigned randomly with one quarter samples of the whole set and annotated online, and then compared the performance with DLS which annotated the same images.

All statistical analyses, including receiver operating characteristic (ROC) curves, were carried out using the programming language Python (version 2.7; Python Software Foundation; Wilmington, DE, USA). The results of the indicators are presented as values with 95% confidence intervals.

Results:

A total of 64914 CFPs were enrolled in this study with the field of 35-55 degrees of the posterior pole covering the whole area of macula and the optic disc. The DLS was trained and validated using 46501 and 10237 images respectively, and evaluated on the three test sets with 3579 images (2635 patients with a mean age (±SD) of 55.4±18.3 ranging from 2 to 96), 1245 images (757 patients with a mean age (±SD) of 48.7±18.0 ranging from 4 to 89) and 3352 images (2558 patients with a mean age (±SD) of 52.6±20.6 ranging from 3 to 97) respectively. The numbers of images in each category of the internal test set were all over 100, which ensured the reliability of the test results. The two external test sets represented a real clinical scenario and the disease distribution of both tertiary medical center and primary hospitals in China over a certain period of time (Table 1). CFPs with more than one label in the training, validation internal test set, external test set A and B were 3202 (6.9%), 488 (4.8%), 334 (9.3%), 70(5.6%) and 217(6.5%), respectively.

Table 1. The sample size of normal fundus and 12 fundus diseases in the five datasets

Label	Development set		Test sets		
	Training set	Validation set	Intern test set	External test set A	External test set B
	N=46501	N=10237	N=3579	N=1245	N=3352
Normal fundus	19146 (41.2)	4315 (9.3)	1053 (29.4)	441 (12.3)	1804 (50.4)
Retinal vein occlusion	3528 (7.6)	967 (2.1)	531 (14.8)	54 (1.5)	123 (3.4)
Referable diabetic retinopathy	2701 (5.8)	642 (1.4)	285 (8.0)	292 (8.2)	388 (10.8)
Pathological myopic retinal degeneration	8243 (17.7)	989 (2.1)	192 (5.4)	84 (2.3)	113 (3.2)
Retinitis pigmentosa	587 (1.3)	137 (0.3)	130 (3.6)	62 (1.7)	38 (1.1)
Retinal detachment	315 (0.7)	88 (0.2)	110 (3.1)	5 (0.1)	14 (0.4)
Epiretinal membrane	2403 (5.2)	544 (1.2)	268 (7.5)	36 (1.0)	165 (4.6)
Dry age-related macular degeneration	2669 (5.7)	808 (1.7)	267 (7.5)	86 (2.4)	404 (11.3)
Wet age-related macular degeneration	1564 (3.4)	433 (0.9)	146 (4.1)	67 (1.9)	75 (2.1)
Macular hole	266 (0.6)	59 (0.1)	137 (3.8)	1 (0.0)	14 (0.4)
Possible glaucomatous optic neuropathy	3648 (7.8)	544 (1.2)	270 (7.5)	79 (2.2)	227 (6.3)
Papilledema	2882 (6.2)	682 (1.5)	228 (6.4)	78 (2.2)	82 (2.3)
Optic nerve atrophy	1459 (3.1)	462 (1.0)	202 (5.6)	23 (0.6)	150 (4.2)

The results are presented with: number (%).

The model performance on the test sets

We developed a late-fusion multi-label model as well as 12 binary classification models for comparison, and the former achieved a higher mean average precision (mAP) on validation set with statistical significance (P=0.020) (eTables 2, 3). The ROC curves were also listed on line (eFigure2, eFigure3). We therefore selected the late-fusion multi-label model for testing. The threshold of the model on validation set were listed in online materiel (eTable4). The AUCs in the internal test set and the two external test sets were 0.950 (95%CI, 0.942~0.957) to 0.996 (95%CI, 0.994~0.998), 0.931 (95%CI, 0.923~0.939) to 1.000 (95%CI, 0.999~1.000) and 0.934 (95%CI, 0.929~0.938) to 1.000 (95%CI, 0.999~1.000), with corresponding sensitivities of 80.4% (95%CI, 79.1%~81.6%) to 97.3% (95%CI, 96.7%~97.8%), 64.6% (95%CI, 63.0%~66.1%) to 100% (95%CI, 100%~100%), and 68.0% (95%CI, 67.1%~68.9%) to 100% (95%CI, 100%~100%), and corresponding specificities of 89.7% (95%CI, 88.8%~90.7%) to 98.1% (95%CI , 97.7%~98.6%), 78.7% (95%CI, 77.4%~80.0%) to 99.6% (95%CI, 99.4%~99.8%) and 88.1% (95%CI, 87.4%~88.7%) to 98.7% (95%CI,

98.5%~99.0%), respectively. For the major blindness leading diseases, the AUCs of referable DR, possible GON, dry and wet form AMD in the external test sets were 0.965 (95%CI, 0.960~0.971) to 0.986 (95%CI, 0.984~0.988), 0.931 (95%CI, 0.923~0.939) to 0.946 (95%CI, 0.942~0.950), and 0.968 (95%CI, 0.964~0.971) to 0.988 (95%CI, 0.986~0.990), respectively. Table 2 shows the results of the AUC, sensitivity and specificity, of the DLS tested on the three test sets. The ROC curves of the DLS tested in the internal set were as Figure 1 shows. Other ROC results tested in the external sets are listed on the online material. (eFigure4,5)

Table 2. The model's performance on the three test sets.

	Intern test set			External test set A			External test set B		
	Sensitivity	**Specificity**	**AUC**	**Sensitivity**	**Specificity**	**AUC**	**Sensitivity**	**Specificity**	**AUC**
Normal fundus	0.945 (0.938, 0.953)	0.967 (0.961, 0.973)	0.989 (0.985, 0.992)	0.951 (0.945, 0.958)	0.787 (0.774, 0.800)	0.956 (0.950, 0.963)	0.862 (0.855, 0.868)	0.895 (0.889, 0.901)	0.955 (0.951, 0.959)
Referable diabetic retinopathy	0.804 (0.791, 0.816)	0.897 (0.888, 0.907)	0.950 (0.942, 0.957)	0.990 (0.986, 0.993)	0.810 (0.797, 0.823)	0.965 (0.960, 0.971)	0.923 (0.918, 0.928)	0.881 (0.874, 0.887)	0.986 (0.984, 0.988)
Retinal vein occlusion	0.964 (0.958, 0.970)	0.969 (0.963, 0.974)	0.994 (0.992, 0.997)	0.963 (0.957, 0.969)	0.960 (0.953, 0.966)	0.992 (0.990, 0.995)	1.000 (1.000, 1.000)	0.986 (0.983, 0.988)	0.999 (0.998, 0.999)
Pathological myopic retinal degeneration	0.958 (0.952, 0.965)	0.971 (0.965, 0.976)	0.988 (0.984, 0.991)	0.952 (0.945, 0.959)	0.990 (0.986, 0.993)	0.992 (0.989, 0.995)	0.991 (0.989, 0.993)	0.938 (0.934, 0.943)	0.989 (0.988, 0.991)
Retinitis pigmentosa	0.962 (0.955, 0.968)	0.978 (0.973, 0.983)	0.996 (0.994, 0.998)	1.000 (1.000, 1.000)	0.988 (0.985, 0.992)	1.000 (0.999, 1.000)	0.895 (0.889, 0.901)	0.977 (0.974, 0.980)	0.996 (0.995, 0.998)
Retinal detachment	0.973 (0.967, 0.978)	0.981 (0.977, 0.986)	0.996 (0.993, 0.998)	0.800 (0.787, 0.813)	0.981 (0.977, 0.986)	0.992 (0.990, 0.995)	0.786 (0.778, 0.794)	0.987 (0.985, 0.990)	0.992 (0.990, 0.993)
Epiretinal membrane	0.918 (0.909, 0.927)	0.923 (0.915, 0.932)	0.968 (0.963, 0.974)	0.694 (0.679, 0.709)	0.975 (0.970, 0.980)	0.938 (0.931, 0.946)	0.745 (0.737, 0.754)	0.889 (0.883, 0.895)	0.934 (0.929, 0.938)
Dry age-related macular degeneration	0.858 (0.846, 0.869)	0.939 (0.931, 0.947)	0.976 (0.971, 0.981)	0.895 (0.885, 0.905)	0.940 (0.932, 0.947)	0.973 (0.967, 0.978)	0.718 (0.709, 0.727)	0.941 (0.937, 0.946)	0.968 (0.964, 0.971)
Wet age-related macular degeneration	0.842 (0.831, 0.854)	0.953 (0.946, 0.960)	0.964 (0.958, 0.970)	0.925 (0.917, 0.934)	0.894 (0.884, 0.904)	0.974 (0.969, 0.979)	0.920 (0.915, 0.925)	0.971 (0.968, 0.975)	0.988 (0.986, 0.990)
Macular hole	0.876 (0.865, 0.887)	0.963 (0.957, 0.970)	0.978 (0.973, 0.983)	1.000 (1.000, 1.000)	0.978 (0.974, 0.983)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	0.966 (0.962, 0.969)	1.000 (0.999, 1.000)
Possible GON	0.804 (0.791, 0.817)	0.934 (0.925, 0.942)	0.953 (0.946, 0.960)	0.646 (0.630, 0.661)	0.938 (0.930, 0.946)	0.931 (0.923, 0.939)	0.797 (0.790, 0.805)	0.930 (0.925, 0.935)	0.946 (0.942, 0.950)
Papilledema	0.904 (0.894, 0.913)	0.950 (0.943, 0.957)	0.980 (0.975, 0.985)	0.756 (0.742, 0.770)	0.990 (0.986, 0.993)	0.991 (0.989, 0.994)	0.756 (0.748, 0.764)	0.975 (0.972, 0.978)	0.990 (0.988, 0.992)
Optic nerve atrophy	0.950 (0.943, 0.958)	0.946 (0.938, 0.953)	0.989 (0.985, 0.992)	0.826 (0.814, 0.838)	0.996 (0.994, 0.998)	0.996 (0.994, 0.998)	0.680 (0.671, 0.689)	0.952 (0.947, 0.956)	0.955 (0.951, 0.959)

To further understand the model's performance, we used heat maps for visualization and clinical review. Figure 2 shows heat maps of the true-positive reports normal fundus and 12 fundus diseases on the external test sets. Different colors mark subregions with different degrees of activation of the DLS, which increase progressively from blue to red as indicated by the color bar. The heat maps indicate that the features extracted by the model generally present a high consistency with human doctors' diagnostic basis in real clinical work according to the specific lesions on CFPs. Some false-positive and false-negative cases indicated that the DLS seemed to miss some fine abnormalities like the change of the disc rim, optic disc pit in possible GON or small macula hole (Figure 3).

We also noticed that the model achieved a relatively lower sensitivity on the detection of possible GON. To further interpret and prove the model's performance, we compared our DLS with some other specialized GON detecting models using public available dataset. The test was performed on Retinal Fundus Glaucoma Challenge, REFUGE (https://refuge.grand-challenge.org) test set, which contains 400 fundus images with 360 normal fundus and 40 glaucoma. We achieved 0.955 AUC

and 0.931 reference sensitivity, which rank six and four among all the 12 participating team, that is comparable to the state-of-the-art models (reference sensitivity: 0.725~0.976, AUC: 0.846~0.989)[25]. The detailed comparison results were available in online materiel (eTable5, eFigure4).

The comparison between human doctors and the DLS model

The mean sensitivity, specificity of the four human doctors were 69.5%, 75.7%, 74.0% and 71.1%, and 98.1%, 97.8%, 97.8% and 97.6%, respectively. The corresponding DLS model's sensitivity and specificity were 90.2%, 86.8%, 84.0% and 82.4%, and 97.6%, 92.6%, 93.7% and 93.6%, respectively. Statistical analysis (Mann-Whitney U test) showed that the DLS achieve significant higher sensitivity comparing with two of the four doctors and lower specificity comparing with all four doctors. Detailed results are available on online materials (eTable6).

Discussions:

DL models for the detection of multiple fundus diseases

Previous studies have reported a large number of DLSs used for multiclassification,

such as the detection of several diseases or severity of DR and AMD using CFPs or optical coherence topography (OCT) [9,16,26]. There have also been studies focused on the detection of multiple fundus lesions recently[21]. The detection of certain fundus diseases using DLS exceeding 10 categories remains very rare. Choi JY. et al.[27] described automated differentiation between normal fundus and 9 retinal diseases but achieved an accuracy of only 36.7% for all 10 classes. Comparing with their study, our work were carried out using a large data set with over 60000 images acquired from real clinical patients. The DLS developed by Son *et al.*[21] proposed a deep learning method for detecting multiple lesion-level abnormalities in color fundus images. The strength to their study is that the detected lesions provide a more intuitive interpretation than holistic predictions as made by the prior art. However, as there lacks a one-to-one correspondence between lesions and fundus diseases, a gap naturally exists when converting lesion-level findings to diseases, which is left untouched by Son et al. in this work, we take a orthogonal direction, making a novel attempt to directly recognize 12 fundus diseases from a given color fundus image. Moreover, we adopt the Class Activation Mapping (CAM) technique to visualize

which part of the given image is responsible for the final prediction.

Furthermore, the diseases selected in this study mostly comprise leading causes of blindness that need early detection and intervention covering a broad spectrum including retinal vascular diseases (RVO, referable DR), retinal degeneration diseases (PM retinal degeneration, RP, RD), macular disease (ERM, AMD and MH) and optic nerve disorders (possible GON, papilledema and optic nerve atrophy). Most of them have rarely been reported in previous studies.

The development and selection of the models

The models developed for multi-disease detection were diverse in previous studies. The scenario targeted most often by machine learning methods for applications in ophthalmology is image classification[28], which is typically used in retinal analysis for automatic screening. Multi-class classification is used[28] to detect the type of disease present or to accurately determine the stage of disease. This has been done for DR[10,11] and ROP[29,30]. In the case of multi-class classification, images belong to only one of the mutually exclusive categories. Choi J. Y. et al.[27] reported a multi-disease recognition model that applied a method of classification to classify fundus images

into different categories of retinal diseases for diagnosis. The authors attributed part of the dissatisfactory performance of the model to decreased expected accuracy as the number of categories multiplied, which has been demonstrated in previous studies[31]. However, mutually exclusive multi-classification model may not be unsuitable for multiple diseases recognition since some fundus diseases may coexist. For example, patients could have DR and ERM simultaneously[32], and the incidence rate of open angle glaucoma in patients with RVO is significantly higher than that in the general population[33]. Our multilabel model was developed with the modified feature layer of SeResNext50 in order to simultaneously classify abnormal versus normal CFP images, and to accurately detect the presence of multiple diseases. We combined the two steps into a single model to simplify implementation in future clinical practice.

The data sets and the model's performance

Our model was trained and tested in real clinical data sets, and this was an important feature of the study, mimicking real screening scenarios as closely as possible at this early stage of development. To assure the accuracy, diversity and reliability of the data sets, we used CFPs from real-life data sets from three different clinical centers

that were annotated by 17 experienced ophthalmologists. The amount of work involved in annotating the images was formidable, and this data set was much larger than in previous studies on multi-disease classification with only 279 images[27]. To our knowledge, this is also the largest multi-disease recognition data set thus far.

Considering the future application scenarios of the model is screening especially in lower level medical places, which maybe accompanied with more complex conditions and interferences while screening, we provided two external test sets from tertiary medical center and primary hospitals respectively. The results showed that the diseases distribution was different from that of tertiary hospital. For example, the proportion of dry AMD and possible GON were much higher. Even so, the results still supported, that the DLS could do well in both scenarios, which proved the possibility of large-scale screening in the future work.

Notice that for glaucoma detection, the sensitivity of our DLS varies, which is 0.913, 0.797 and 0.646 on REFUGE, the external test set B and the external test set A, respectively. We attribute this variation to the distinct sources of the three test sets. REFUGE, as a public benchmark dataset, tends to include images of less ambiguity to

ensure the reliability of its ground truth. Indeed, we observed that images from this dataset are typical with respect to glaucoma. Recall that the external test set B and A were collected from primary hospitals and tertiary hospitals respectively. Given the common practice of a referral medical system, where cases that are less typical and thus more difficult to diagnose are to be referred from a primary hospital to a tertiary hospital, it is fair to claim that images from A were the most challenging. The increasing difficulty in glaucoma diagnosis from REFUGE to the test set B and to the test set A explains the decreasing sensitivity of the DLS to detect this condition.

The interpretation of the heat maps

The “black box” problem of DLS has greatly limited its application and acceptance in real clinical practice. In this study, we used heat maps for visualization. As the heat maps indicated, the features extracted by the model for prediction are very similar to human doctors’ considerations. Taking referable DR as an example (Figure 2, O), the model precisely extracted the appropriate retinal lesions (intraretinal and preretinal hemorrhages) and provided a correct prediction. The heatmaps are also helpful on understanding the false results. For example, the heatmap indicted that in false

negative case of possible GON (Figure 3 A2), the model payed almost no attention on the optic disc and failed to give the correct answer. The DLS model presented a limited performance on the detection of specific diseases like possible GON. To further interpret the results, we tested the model in a public available

Limitations and future works

Our work has some limitations. Firstly, while we have spent much efforts to expand our external test sets, the testing sample sizes for MH and RD, which are 19 and 15 in total, remain relatively small, as compared to the other conditions. To improve the reliability of the detection performance of the two diseases, more test samples need to be collected for future exploration. Secondly, the external evaluation on a clinical dataset collected from tertiary hospitals (external test set A) shows that our DLS detects glaucoma with a relatively lower sensitivity of 0.646. Given that glaucoma is a major blinding disease, much work remains to be done for real-world deployment. Thirdly, some diseases included in this study initiate from the peripheral retinal area such as RP and RD, but most of the images we used for analysis were centered by the macula fovea with the maximal field of 55 degree. Therefore, the

detection of these diseases may be limited. With the future common use of ultrawide fundus camera, DLS model for this kind of CFP is of high research value. Finally, future prospective trials are needed to assess the DLS in multiple independent real clinical scenarios.

Conclusion:

The proposed DLS showed well performance on the three test sets for the detection of normal fundus as well as 12 major fundus diseases. The application of this model may alleviate the workloads of trained specialists and provide an efficient, low-cost approach for preliminary screening in places with scarce medical resources and ophthalmologists. Further acquisition of data to broaden the extent of screening for more fundus diseases will be the next step of our work.

Acknowledgments

The authors thank Di Gong, Hong Du, Ning Chen, Dongmei Huo, Nan Chen, Hongling Chen, Donghui Li, Meiyan Zhu, Yanting Wang, Xiao Chen, Hui Liu, Huan

Chen and Tong Zhao for their valuable contribution to this research. They devoted considerable time and effort to this work during the process of online annotation that lasted for more than 8 months.

Funding statement:

This work was supported by:

1. Chinese Academy of Medical Sciences (CAMS) Initiative for Innovative Medicine (CAMS-12M), grant number: 2018-I2M-AI-001

2. Pharmaceutical collaborative innovation research project of Beijing Science and Technology Commission, grant number: Z191100007719002

3. Beijing Natural Science Foundation Haidian original innovation joint fund, grant number: 19L2062

4. Beijing Natural Science Foundation, grant number: 4202033

5. The priming scientific research foundation for the junior researcher in Beijing Tongren Hospital, Capital Medical University, grant number: 2018-YJJ-ZZL-052

Competing Interests Statement:

All authors declare that No conflict of interest exists.

Contributorship Statement:

Bing Li, contributed to the statistical analysis, drafting and revising of the manuscript.

Huan Chen, Bilei Zhang and Mingzhen Yuan contributed to the standard operating procedure and quality control of the datasets.

Xuemin Jin, Bo Lei, Jie Xu, and Wei Gu contributed to the acquisition of the color fundus photograph of the datasets.

David Chuen Soong Wong contributed to the revision of the manuscript.

Xixi He and Hao Wang contributed to the models' developing, statistical analysis and preparing of the figures for the work.

Xirong Li and Dayong Ding contributed to the development of the models and interpretation of data, and revision of the manuscript for this study.

Youxin Chen and Weihong Yu contributed to the conception and design of the work, revision of the manuscript and will final approval of the version to be published.

References

[1] Quigley HA, Broman AT. The number of people with glaucoma worldwide in 2010 and 2020. *Br J Ophthalmol* 2006;90(3):262-7.

[2] Wong WL, Su X, Li X, et al. Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. *Lancet Glob Health* 2014;2(2):e106-16.

[3] Guariguata L, Whiting DR, Hambleton I, et al. Global estimates of diabetes prevalence for 2013 and projections for 2035. *Diabetes Res Clin Pract* 2014;103(2):137-49.

[4] Yau JW, Rogers SL, Kawasaki R, et al. Global prevalence and major risk factors of diabetic retinopathy. *Diabetes Care* 2012;35(3):556-64.

[5] Shaw JE, Sicree RA, Zimmet PZ. Global estimates of the prevalence of diabetes for 2010 and 2030. *Diabetes Res Clin Pract* 2010;87(1):4-14.

[6] Feng JJ, An L, Wang ZF, et al. [Analysis on ophthalmic human resource allocation and service delivery at county level in Mainland China in 2014]. *Zhonghua Yan Ke Za Zhi* 2018;54(12):929-34.

[7] Gao X, Lin S, Wong TY. Automatic Feature Learning to Grade Nuclear Cataracts Based on Deep Learning. *IEEE Trans Biomed Eng* 2015;62(11):2693-701.

[8] Caixinha M, Amaro J, Santos M, et al. In-Vivo Automatic Nuclear Cataract Detection and Classification in an Animal Model by Ultrasounds. *IEEE Trans Biomed Eng* 2016;63(11):2326-35.

[9] Ting DSW, Cheung CY, Lim G, et al. Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. *JAMA* 2017;318(22):2211-23.

[10] Li Z, Keel S, Liu C, et al. An Automated Grading System for Detection of Vision-Threatening Referable Diabetic Retinopathy on the Basis of Color Fundus Photographs. *Diabetes Care* 2018;41(12):2509-16.

[11] Gulshan V, Peng L, Coram M, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* 2016;316(22):2402-10.

[12] Li Z, He Y, Keel S, et al. Efficacy of a Deep Learning System for Detecting Glaucomatous Optic Neuropathy Based on Color Fundus Photographs. *Ophthalmology* 2018;125(8):1199-206.

[13] Redd TK, Campbell JP, Brown JM, et al. Evaluation of a deep learning image assessment system for detecting severe retinopathy of prematurity. *The British journal of ophthalmology* 2018;

[14] Brown JM, Campbell JP, Beers A, et al. Automated Diagnosis of Plus Disease in Retinopathy of Prematurity Using Deep Convolutional Neural Networks. *JAMA ophthalmology* 2018;136(7):803-10.

[15] Li F, Chen H, Liu Z, et al. Fully automated detection of retinal disorders by image-based deep learning. *Graefes Arch Clin Exp Ophthalmol* 2019;257(3):495-505.

[16] Grassmann F, Mengelkamp J, Brandl C, et al. A Deep Learning Algorithm for Prediction of Age-Related Eye Disease Study Severity Scale for Age-Related Macular Degeneration from Color Fundus Photography. *Ophthalmology* 2018;125(9):1410-20.

[17] Loo J, Fang L, Cunefare D, et al. Deep longitudinal transfer learning-based automatic segmentation of photoreceptor ellipsoid zone defects on optical coherence tomography images of macular telangiectasia type 2. *Biomed Opt Express* 2018;9(6):2681-98.

[18] Kihara Y, Heeren TFC, Lee CS, et al. Estimating Retinal Sensitivity Using Optical Coherence Tomography With Deep-Learning Algorithms in Macular Telangiectasia Type 2. *JAMA Netw Open* 2019;2(2):e188029.

[19] Lu W, Tong Y, Yu Y, et al. Deep Learning-Based Automated Classification of Multi-Categorical Abnormalities From Optical Coherence Tomography Images. *Transl Vis Sci Technol* 2018;7(6):41.

[20] Liu YY, Ishikawa H, Chen M, et al. Computerized macular pathology diagnosis in spectral domain optical coherence tomography scans based on multiscale texture and shape features. *Invest Ophthalmol Vis Sci* 2011;52(11):8316-22.

[21] Son J, Shin JY, Kim HD, et al. Development and Validation of Deep Learning Models for Screening Multiple Abnormal Findings in Retinal Fundus Images. *Ophthalmology* 2019;

[22] Ferris FL, 3rd, Wilkinson CP, Bird A, et al. Clinical classification of age-related macular degeneration. *Ophthalmology* 2013;120(4):844-51.

[23] Hu J, Shen L, Albanie S, et al. Squeeze-and-Excitation Networks. *IEEE Trans Pattern Anal Mach Intell* 2020;42(8):2011-23.

[24] Zhou B, Khosla A, Lapedriza A, et al. Learning Deep Features for Discriminative Localization. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA, 2016:2921-29.

[25] Orlando JI, Fu H, Barbosa Breda J, et al. REFUGE Challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Med Image Anal* 2020;59:101570.

[26] Keel S, Wu J, Lee PY, et al. Visualizing Deep Learning Models for the Detection of Referable Diabetic Retinopathy and Glaucoma. *JAMA Ophthalmol* 2019;137(3):288-92.

[27] Choi JY, Yoo TK, Seo JG, et al. Multi-categorical deep learning neural network to classify retinal images: A pilot study employing small database. *PLoS One* 2017;12(11):e0187336.

[28] Schmidt-Erfurth U, Sadeghipour A, Gerendas BS, et al. Artificial intelligence in retina. *Prog Retin Eye Res* 2018;67:1-29.

[29] Zhang YS, Wang L, Wu ZQ, et al. Development of an Automated Screening System for Retinopathy of Prematurity Using a Deep Neural Network for Wide-Angle Retinal Images. *Ieee Access* 2019;7:10232-41.

[30] Wang JY, Ju R, Chen YY, et al. Automated retinopathy of prematurity screening using deep neural networks. *Ebiomedicine* 2018;35:361-68.

[31] Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542(7639):115-18.

[32] Jackson TL, Nicod E, Angelis A, et al. Vitreous attachment in age-related macular degeneration, diabetic macular edema, and retinal vein occlusion: a systematic review and metaanalysis. *Retina* 2013;33(6):1099-108.

[33] Na KI, Jeoung JW, Kim YK, et al. Incidence of Open-angle Glaucoma in Newly Diagnosed Retinal Vein Occlusion: A Nationwide Population-based Study. *J Glaucoma* 2019;28(2):111-18.

Figure legends

Figure 1 The receiver operating characteristic (ROC) curves of the DLS tested in the internal test set.

Figure 2 CFPs and visualization heat maps of true-positive cases on the internal test set. The color bar marks subregions with different active intensities of the model, which increase progressively from the blue end to the red end. These heat maps represent the ability of our method to objectively distinguish different diseases.

Figure 3. The fundus image and corresponding heat maps of some cases of false positive and false negative results predicted by the DLS in the validation set. A1 and A2 are false negative cases: the DLS miss diagnosed referable diabetic retinopathy (A1) and possible GON (A2) to normal fundus; B1 and B2 are false positive cases: the DLS miss diagnosed macular hole to wet age-related macular degeneration.